Football Match Classification Project Protocol

Project Overview

This project was completed as part of the Machine Learning course at Bar-Ilan University.

It involves a full data science pipeline applied to football match data with the goal of building a multi-class classification model to predict match results (home win, draw, away win).

Tools & Environment

The project was developed in Jupyter Notebook using Python.

Key libraries include:

- pandas, numpy, matplotlib, seaborn
- scikit-learn, xgboost
- autoviz for EDA
- mplsoccer for football-specific stats

Workflow

1. **Data Preprocessing**: Loading data, fixing datatypes, creating datetime features
2. **EDA (Autoviz & Manual)**: Visualization and summary stats
3. **Handling Outliers & Missing Values**
4. **Feature Engineering**: Cumulative stats, rolling averages, ratios, interaction terms
5. **Feature Selection**: Using univariate tests, LASSO, tree-based models
6. **Modeling & Hyperparameter Tuning**: Gradient Boosting, Random Forest, SVM, Logistic Regression, XGBoost
7. **Model Evaluation**: Accuracy, F1, Precision, Recall, Log-loss, AUC

Football Match Classification Project Protocol

Final Model

The best-performing model was an XGBoost classifier with the following configuration:

- learning_rate=0.05

- max_depth=110

- min_child_weight=50

- n_estimators=400

- eval_metric='logloss'

This model achieved over 99% accuracy, precision, and recall on test and dev sets.